evaluation

Anastasia.Bezerianos@lri.fr

does it work?

why evaluate?

Initial design phases
- Develop and evaluate initial design ideas with users (participatory design)

Iterative design
- System behavior corresponds to user needs
- Solve specific problems
- Choice between alternatives

Acceptance testing
- Verify that the system addresses the user needs

Ideal: evaluate with real user populations

Evaluation Techniques

Informal and quick:
- Heuristics
- Heuristic evaluation
- Design Walkthrough
- Others …

Formal and targeted:
- Alternatives User Studies
- Controlled Experiments
- Quasi-experiments
- Others (Interviews, Questionnaires, Observations)

evaluation:
informal and quick

Design Walkthrough

A group evaluates an aspect of a specific "something" step-by-step:

program source code	to find bugs
system architecture design	to understand structure
UI screens	to get user feedback
text (e.g. scientific articles)	to verify its structure and understandability
experiment	to verify the method and details

Design Walkthrough

Goal:

Aid to informally and quickly identify problems, using evaluation criteria (to be defined by you in advance)

Procedure

Choose a small group with different expertise and roles
Fix the duration to 1h max
A presenter describes a scenario (storyboard, video prototype, system)
Choose levels of critiques
The group identifies as many problems as possible
Use rules to aid in problem finding
(e.g. design principles, specifications, usability criteria, task sequence)

Design Walkthrough : Types of comments

Specific

e.g. it needs 3 steps to do a simple search

Missing Functions

e.g. no help provided, need search widget

Bugs

e.g. the import functionality does not work

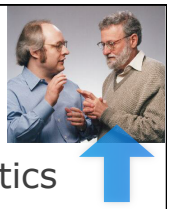
Suggestions

e.g. provide an overview of the data generated

General (the least useful)

e.g. difficult to use, too many icons

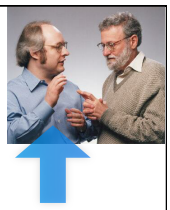
Heuristics - Norman (1983)



You can use the design principles as heuristics for testing:

1. Visibility: state of the system observed in the UI
2. Affordances: perceived actions
3. Mapping: correspondence between action and result
4. Feedback (and Feedforward): inform the user
5. Metaphors and negative transfers
6. Constraints: use to avoid errors

Heuristic Evaluation - Nielsen (1990)



More formal than heuristics but quick

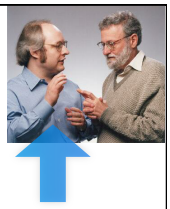
Systematic inspection of the interface, using usability categories

Process

- 3-5 inspectors (usability experts, end-users)
- Inspect the interface (approx. 1-2 hours for simple interfaces)
- Compare their notes afterwards

Works for storyboards, prototypes, real systems (can even do it yourself)

Heuristic Evaluation - Nielsen (1990)



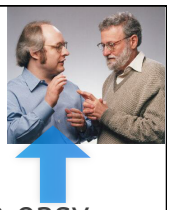
Open, non-guided

- Exploration of system without specific task
- Helps explore different aspects of interface

Guided by Scenarios

- Use representative user tasks or scenarios
- Problems identified in problematic parts of the system
- Evaluate functions of interest
- … but problems can be missed

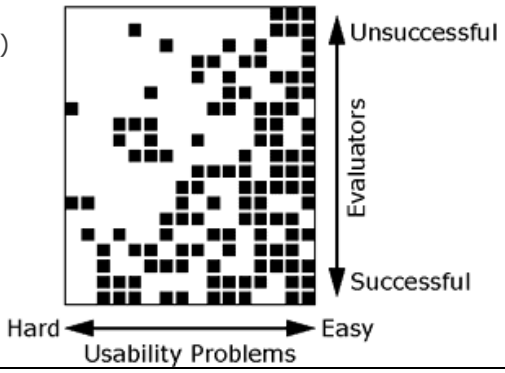
Heuristic Evaluation - Nielsen (1990)



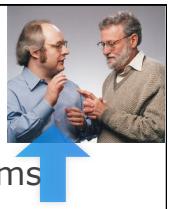
Evaluators/inspectors can miss problems (both easy and hard to find)

"Best" evaluators can miss easy problems
"Bad" evaluators can discover difficult problems

Example of an evaluation)

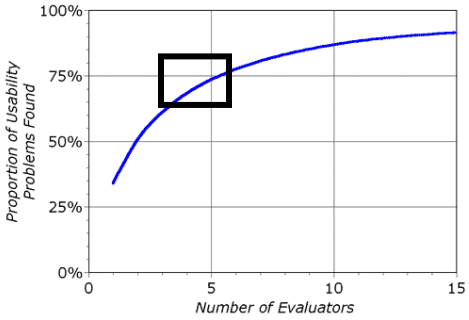


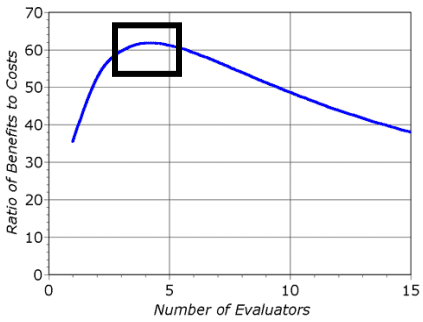
Heuristic Evaluation - Nielsen (1990)



3-5 evaluators find 66-75% of usability problems

different evaluators find different problems if they work independently from each other





Evaluation Techniques

Informal and quick: possible at different stages in the cycle

Heuristics :
- you or experts
- tests usability

Heuristic Evaluation
- evaluators, experts or you
- tests usability mostly (especially Open Evaluation)

Design Walkthrough
- evaluators, experts
- utility, usability (depending on the criteria used)

evaluation:
formal and targeted

Others: we already know

Some formal and lengthy:
Interviews, Questionnaires, Observations

What we learned in "understanding users":
choice of questions (Interviews & Questionnaires)
avoid influencing users (all)

and analysis done using the same methods
e.g. grounded theory or statistics (a bit on this next)

Alternatives (or Usability Studies)

Usability Study (not the same as heuristic eval.)
Test alternatives for the system with users

e.g.

interaction techniques	pallet vs. menus
icon organization	list vs. array
help	tutorials vs FAQ
design alternatives ...	

Usability Study

Goal: Determine best design choice, with users

Procedure:

Describe the purpose of the design (and alternatives)
Choose the *dependent* & *independent* variables to test
(what you measure & what you compare)
Make a prediction/hypothesis
Prepare the environment for each test condition (alternative)
Use at least 3 subjects (5 better)

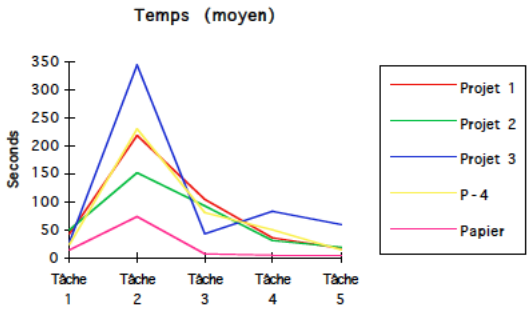
Analyze the results
Are the differences significant?

Usability Study

Dependent variables (measures):
what you measure

Usually tested in HCI :
- Efficiency
- Errors
- Satisfaction
- Learnability
- Memorability

Example of results: Time



usually accompanied by a report of identified (usability) problems

Controlled Experiments

Goal: Does the treatment X cause the effect Y?

Usually interested in comparing with X, without X

More formal than usability studies

Hypothesis testing

Compare alternative hypotheses

Control conditions to isolate the variables you want to test

Analysis of correlations or differences

Measure the degree of correlation between two factors

Knowing one helps predict the other

Examine if there is a difference between two factors

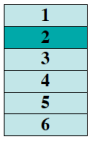
Y is affected differently under treatment X

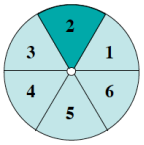
Design a simple Controlled Experiment

1. Specify the hypothesis
   What do we compare and what do we predict?
2. Specify the independent variables
   What changes?
3. Operationalize the behavior (remove biases and noise)
   What are we studying?
4. Specify the dependent variables
   What are we measuring?
5. Specify procedures
   What are the experimental and control groups?
6. Identify the appropriate statistical tests
   Is there a difference?

Example of a hypothesis

Compare linear to pie menus





Hypothesis: pie menus are faster

Null hypothesis (that we will try to disprove with statistics):

There is no difference in user performance in terms of time and error rate for the selection of an item in a linear and in a pie menu, regardless of previous user experience of using a mouse or other types of menus.

Variables: Independent / Dependent

Independent variables (= factors) are those we want to verify or that we want to control, independently of each other

e.g.
2 Types of menus : linear, pie
5 Number of menu items : 3, 6, 9, 12, 15
3 levels of experience : expert, novice, intermediate
=> 2 x 5 x 3 = 30 unique *conditions*

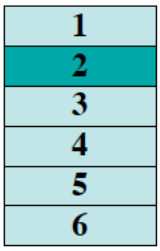
Dependent variables (= measures) are those we measure, they depend on the behavior of the subject and (hopefully) the independent variables

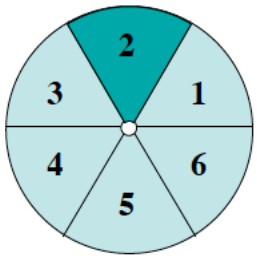
e.g. in HCI
Time to select an item
Number of errors
Others?

For statistical analysis we need adequate measures (user data) per condition

Operationalize the behavior

In our experiment:
- Same labels for menu items
- Same menu position (center of screen)
- View the item to select instead of having to find it





Statistics

This is a VERY large domain

It is difficult to make correct assumptions

Errors are common

You can do simple tests

Statistical analysis

Provide the mathematical characteristics of data
Describes how data sets are related
Estimates the probability that hypothesis are correct

Descriptive Statistics:
Reduce amount of data: e.g.: mean, distribution

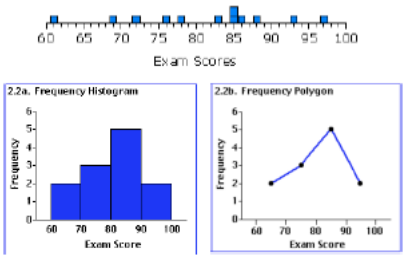
Inferential Statistics:
Infer population properties from a small sample
e.g.: measure the probability than an observed difference is real

Descriptive Statistics

Simpler but less powerful

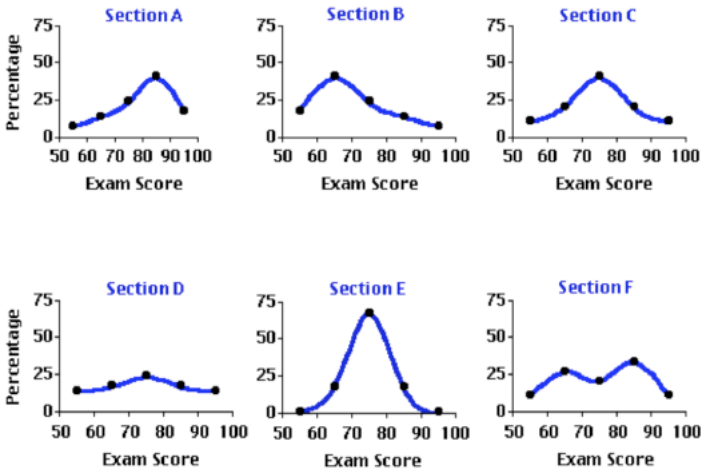
How to summarize a set of measures of a variable

Distribution of frequencies values



Types of distributions
Measures of central tendency
Measures of variability
Measurement of the correlation between two variables

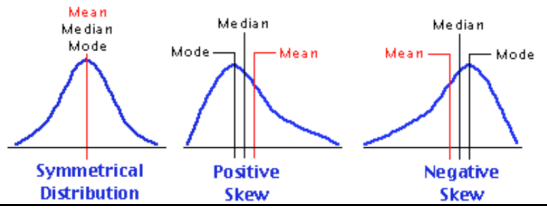
Types of distributions



Measures of central tendency

Applicable to scalar variables

Average/Mean:	Sum of values divided by their number
Median:	"middle" value of the N sorted values N odd: index value (N +1) / 2 N even: average index values N/2, N/2+1
Mode:	the most frequent value There may be several modes (e.g. 2 modes = bimodal)



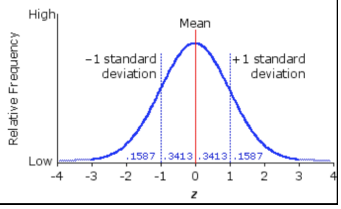
Mesures of variability

Measure the "spread" of the distribution

Range: distance between min and max value

Variance and standard deviation:

sum of squares between each value and the mean
variance: $s2 = \sum (Xi - M)2 / N$
standard deviation: $s = sqrt (\sum (Xi - M)2 / N)$



Correlation between two variables

Measures the relation between two scalar variables

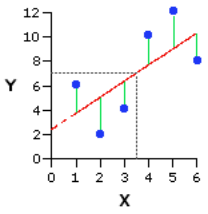
In general an independent variable X and a dependant Y

Coefficient of linear correlation $r (-1 \leq r \leq 1)$

$r = \sum (Xi - MX)(Yi - MY) / sqrt (\sum(Xi - MX)2 \sum(Yi - MY)2)$

r2 can be interpreted as the portion of variable Y associated to X
1- r2 is the residual variance (what cannot be explained)

ATTENTION:
correlation does not imply **cause**



Inferential Statistics

Complex, more powerful than descriptive statistics

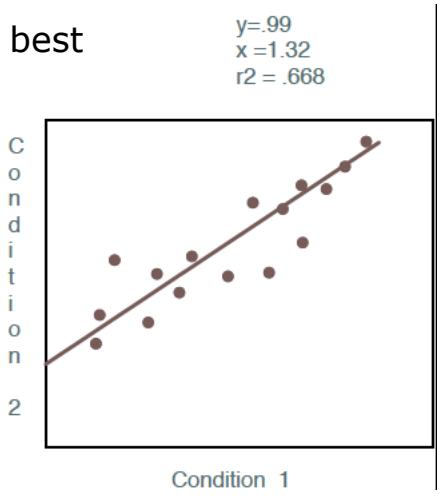
Based on probability theory

E.g.: Comparing Means
Student test (t-test), ANOVA
E.g.: correlation
Pearson rho factor
E.g.: Regression Analysis

Regression

Calculates the line that best approximates the data

Use one variable to predict another



Using statistical methods

Ensure that the statistical test is valid, based on

Population distribution	(e.g. normal)
Data type	(e.g. ordinal)
Sampling procedure	(e.g. random)
Sample size	(e.g. n=30, close to normal)

Determine the degree of confidence of your results
"The assumption that prior experience of using the mouse makes no difference is rejected with a p **level of 0.05** "

Interpret your results

Statistical Significance

P-value: A criterion alpha = 0,05

0,05 = 1/20

If there is no difference and I did this experiment 20 times, one test will give a significant result if it is not true, the other 19 will produce a non-significant result
(5% chance of random observation)

Signification statistique

Provides a quantitative estimate of the probability that two distributions are different

If the number of subjects is large, a small difference can produce a significant result

If you don't have enough data you may not be able to see a significant effect that exists, or see one that does not

And very importantly:

significance ≠ importance